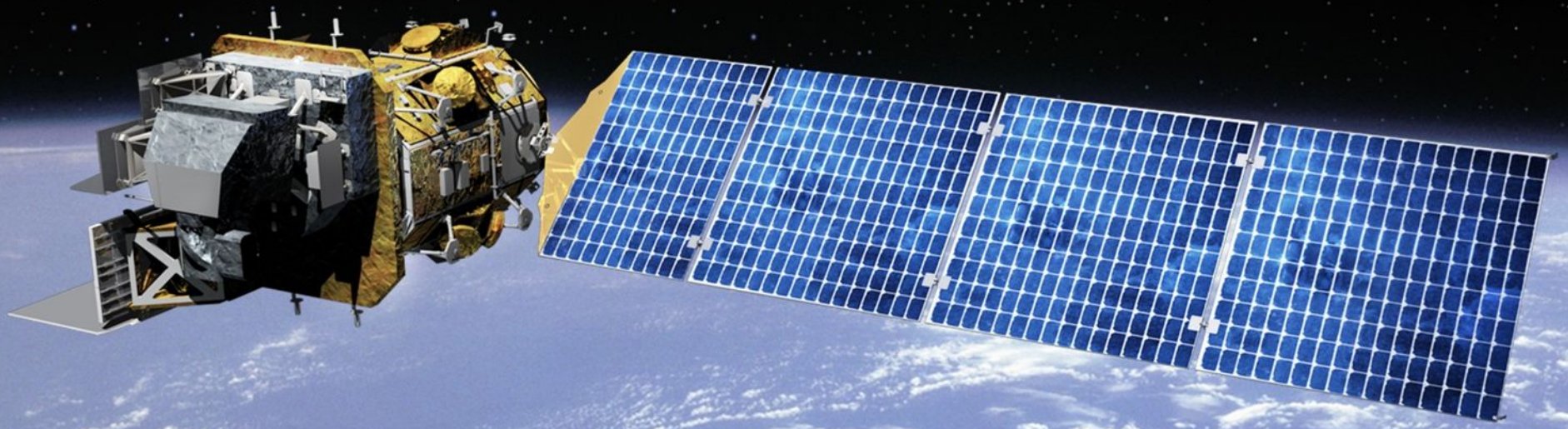


Thoughts on cloud migration

Pete Doucette
U.S. Geological Survey
Department of the Interior

Cloud Summit
NOAA / NESDIS
November 21, 2019

USGS
science for a changing world

Cloud motivations— price vs. value

Oscar Wilde, "Lady Windermere's Fan", 1891

Cecil Graham: What is a *cynic*?
Lord Darlington: A man who knows the **price** of everything, and the **value** of nothing.

Cecil Graham: And a *sentimentalist*, my dear Darlington, is a man who sees an absurd **value** in everything and doesn't know the market **price** of any single thing."

Cloud motivations— price vs. value

1. **Price** (cynic)
    - satisfy operational *requirements*
    - economies of scale can save $

2. **Value** (sentimentalist)
    - data □ knowledge (“*inference*”)
    - speed, agility, flexibility, diversity
    - unlimited scaling

United States Geological Survey

- ~8500 employees
- 65 science centers
- 400 field offices
- National Water Model
- 1000s of streamgages and seismic monitor stations
- National Map and LIDAR
- 2 operational satellites

Motto: Science for a changing world.


Energy and Minerals


Ecosystems

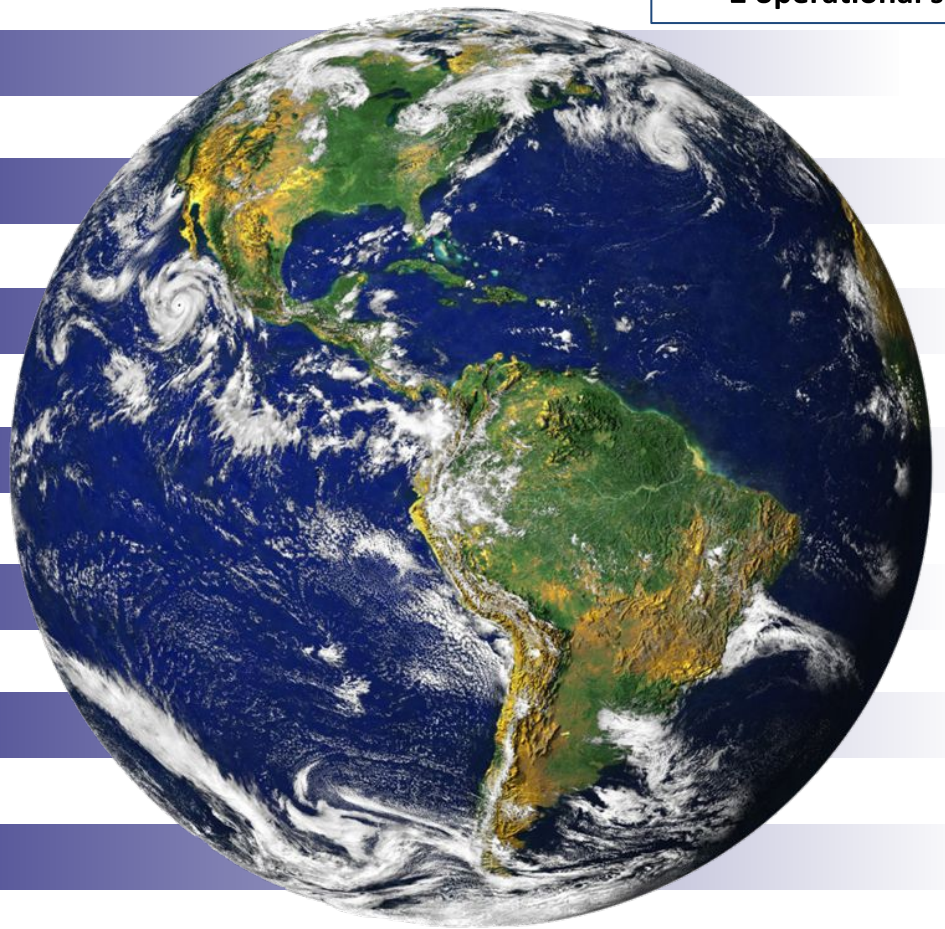

Natural Hazards


Water Resources


Land Resources

Core Science Systems

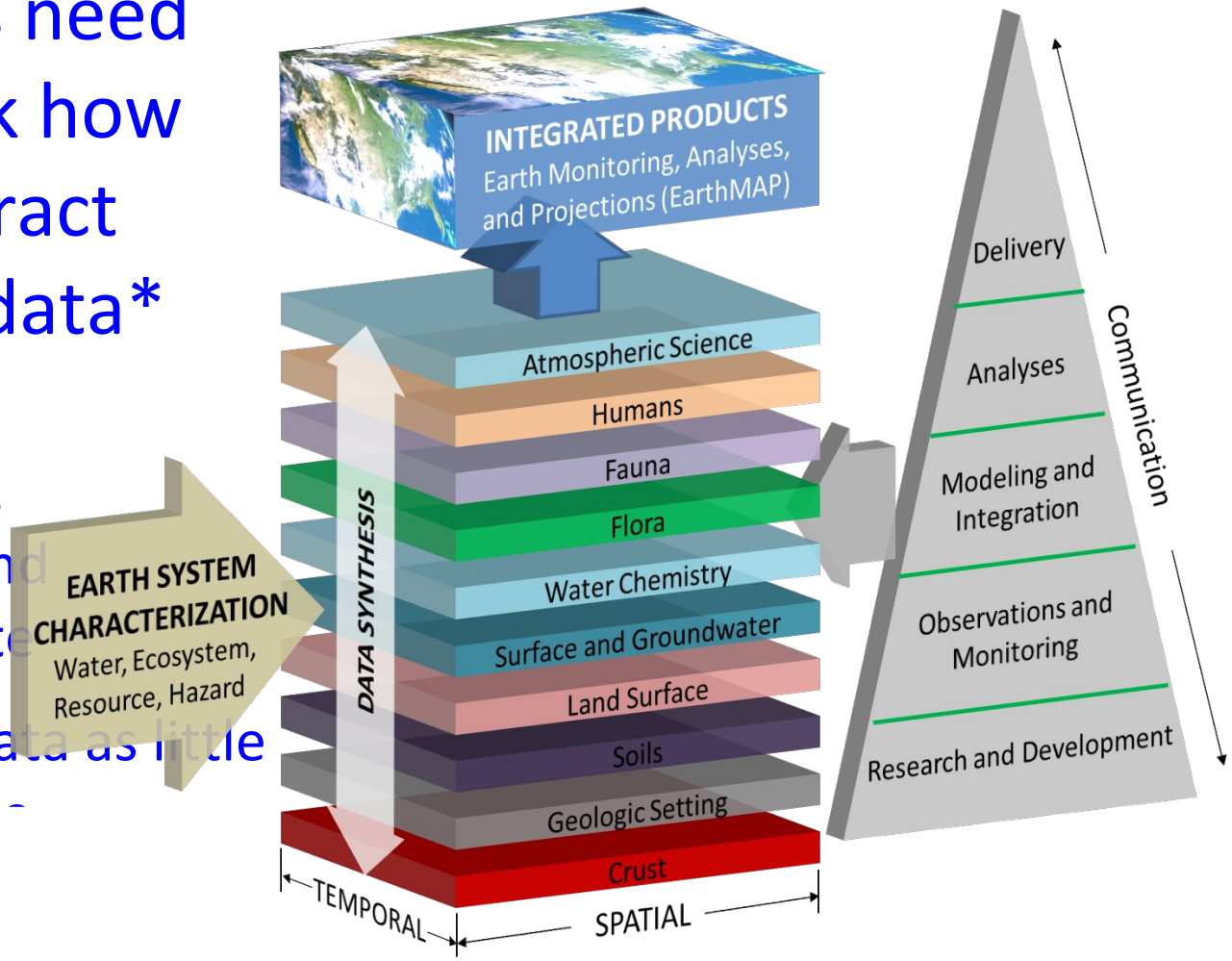

Environmental Health



EarthMAP: The grand challenge

Scientists need to rethink how they interact with big data*

- storage v. on-demand recompute
- moving data as little as possible
- Federated platforms



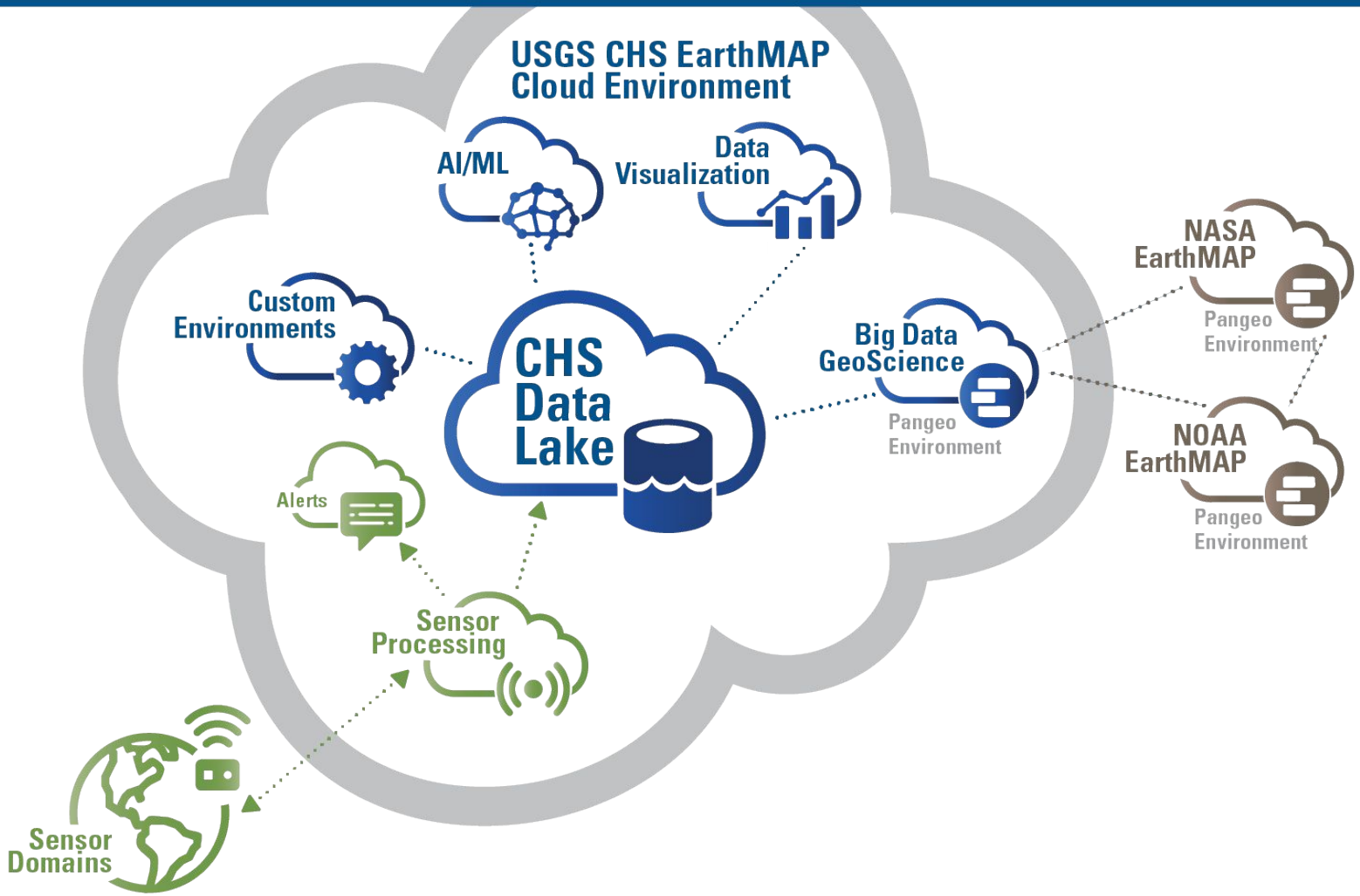
*"Science needs to rethink how it interacts with big data: Five principles for effective scientific big data systems," Niall H. Robinson, Joe Hamman, and Ryan Abernathy (https://arxiv.org/pdf/1908.03356.pdf)

USGS

Graphic source: "**Grand Challenges for Integrated U.S. Geological Survey Science—A Workshop Report," USGS Open-File Report 2017–1076.**

USGS Cloud Hosting Solutions EarthMAP Platform

Simplified Conceptual Phase 1 Architecture



Lessons from USGS cloud migration

- There is no replacement for having Cloud strategy **support from the head of the agency** in clear terms communicated to all.

- Required staffing/**funding** to support a VDC should be **viewed no differently** than support for on-prem data center. Needed skill sets will shift accordingly.

- All things being equal, consider using tools, services, and policies "**Born in the Cloud**" to support the cloud environment.

- A fundamental decision--- Use cloud as an **augmentation OR replacement platform?** Each has its own set of challenges to materialize and support.

Lessons from USGS cloud migration

- More efficient when performed as **many smaller migrations**
  - Big lift-and-shift efforts are rarely seamless
    - Understanding **dependencies** is critical
      - USGS provided access to Redshift Spectrum, but underlying dependence on AWS Glue resulted in inability to use Redshift Spectrum Push to get appropriate tools Fedramp certified.
      - Complicates multi-vendor cloud deployments
- Addressing **cultural** perceptions
  - Cloud is not a fad [OMB-- Cloud First (2010) Cloud Smart (2018)]
  - **Concerns** jobs, funding deficiency/mystery bills, security, vendor lock-in, contract changes
  - Early **on-ramps** for workforce is important
  - Embrace (new) **DevOps** mindset – to some degree, what used to be infrastructure will become code.

PANGEO

A community platform for
Big Data geoscience collaboration

Motivation

- *Big Data*: datasets are growing too rapidly and legacy software tools for scientific analysis can't handle them.

- *Technology Gap*: a growing gap between the technological sophistication of industry solutions (high) and scientific software (low).

- *Reproducibility*: a fragmentation of software tools and environments renders most geoscience research effectively unreproducible and prone to failure.

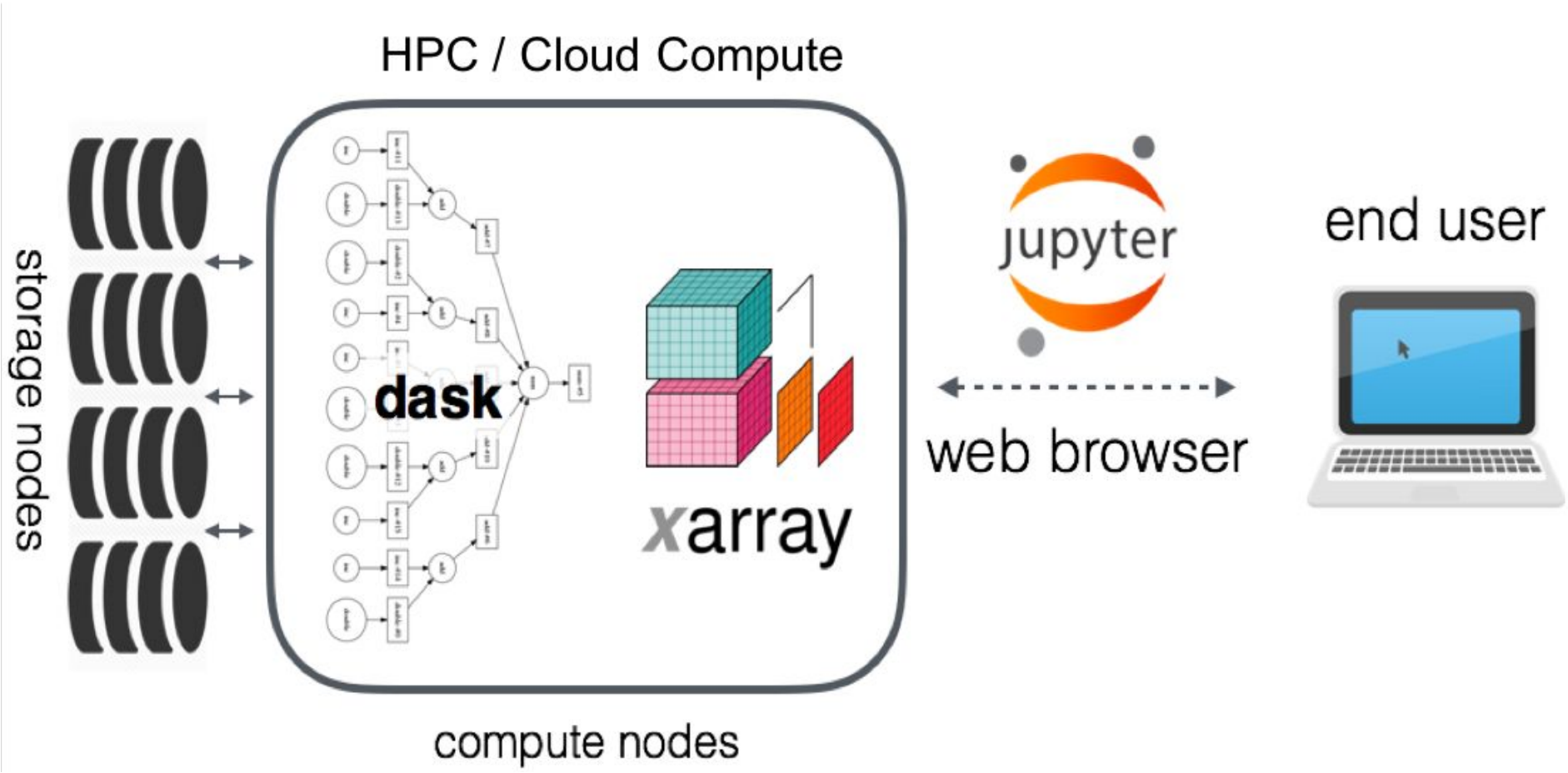
Pangeo aims to address these challenges through a unified, collaborative effort.

Funding sources: NSF EarthCube Transforming Geosciences Research NASA Alfred P. Sloan Foundation

USGS Source: http://pangeo.io

PANGEO

Deployment architecture



USGS

Source: http://pangeo.io/architecture.html

PANGEO

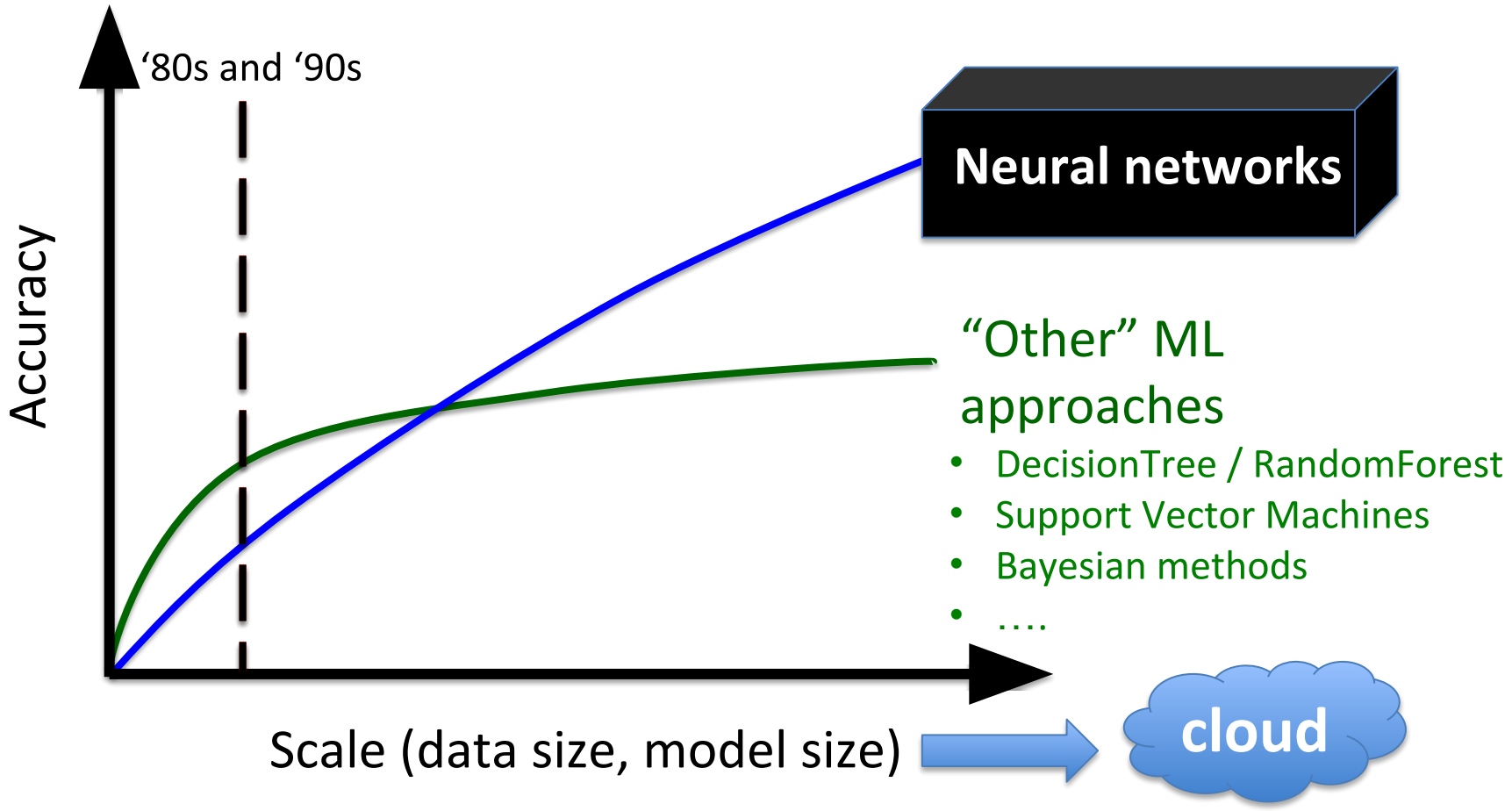
INTERCHANGEABLE PIECES IN PANGEO (PICK 1 OR MORE FROM EACH ROW)

Data Models	xarray	Iris	pandas
N-D Arrays	NumPy	DASK	
Processing Mode	Interactive jupyter	Batch 	Serverless
Compute Platform	HPC 	aws	Google Cloud Platform
Foundation	python™		

USGS

Source: http://pangeo.io/architecture.html

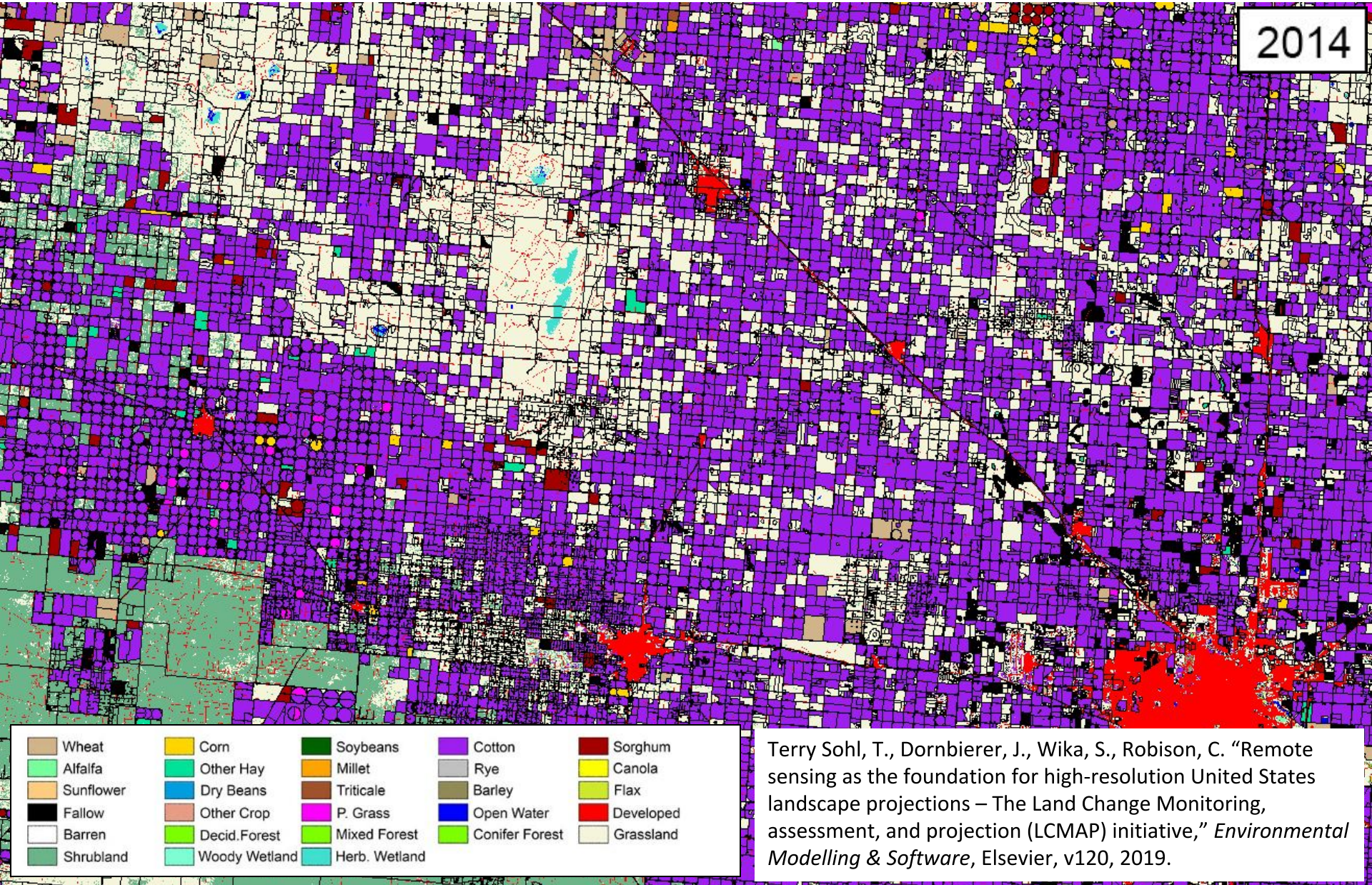
AI/ML (deep learning) trend □ big data



Graph adapted from: https://www.scribd.com/document/355752799/Jeff-Dean-s-Lecture-for-YC-AI

Data science goals – scenario projections

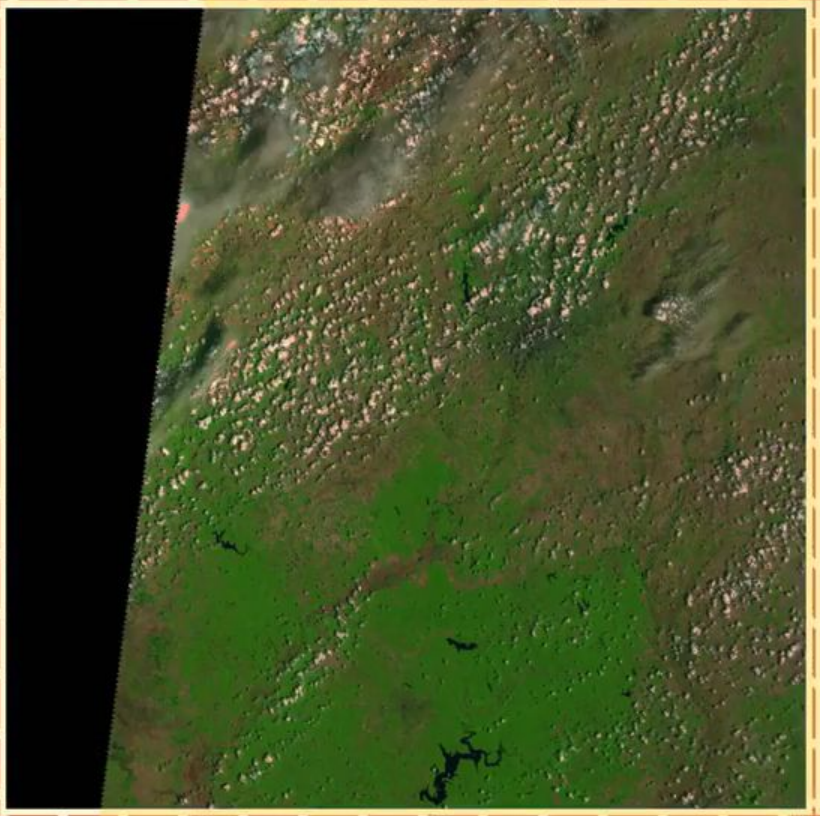
Effects of Ogallala aquifer decline near Lubbock, TX



Terry Sohl, T., Dornbierer, J., Wika, S., Robison, C. "Remote sensing as the foundation for high-resolution United States landscape projections – The Land Change Monitoring, assessment, and projection (LCMAP) initiative," *Environmental Modelling & Software*, Elsevier, v120, 2019.

Backup

Landsat Analysis Ready Data* (ARD)



Currently GeoTIFF bundled as .tar

USGS

*Dwyer, J., Roy, D., Sauer, B., et al.. Analysis Ready Data: Enabling Analysis of the Landsat Archive. *Remote Sens*. 10(9), 1363, 2018.

Cloud Optimized GeoTIFF (COG)

- Customized range requests
- Make a connection from pixels on the screen to specific scenes
- Select different band combinations
- Show change time-lapse